Internship report

Study of AI and Cloud technologies application for enhancement of marketing strategy and user experience

University of Strasbourg

2020

Vadym Hadetskyi

CSMI M2

Table of content

Table of content 2

Introduction 3

The Roadmap 4

1. Context, objectives and evolution of the internship 6

2. Identification of personality profiles 9

1.1 Data 9

1.2 Unsupervised learning and clustering 11

1.3 Dimensionality reduction 11

1.4 Evaluation of the partition 11

3. Chatbot development 12

2.1 The main idea and description of bot's desired capabilities 12

2.2 Dialog structure, entities and intentions 12

4. Cloud Infrastructure 13

3.1 IBM Cloud 13

3.2 Databases 13

3.3 Docker, Kubernetes & Containers 13

Conclusion 14

Bibliography 15

Annex A 16

Introduction

This report aims to present the work conducted in the scope of the internship on the research of possibilities of AI-based methods for marketing strategy and user experience enhancement. The availability of machine learning and cloud technologies opens a lot of ways to improve and optimize numerous processes across various fields of enterprise activity, from production to customer relations. Thus, the goal of the given endeavor is to find out how the company may profit from development and implementation of those techniques.

The internship if offered by Vivialys Groupe — biggest supplier of housing in Grand Est region of France. Main workload is distributed among the projects, that belong to research & development department of the enterprise, which are:

- Development of smart chatbot
- IBM Cloud infrastructure
- Definition of personality profiles
- Automatic configurator of apartment modules

Based on those, principal areas of work and technologies involved in the frame of the internship are:

- Unsupervised learning, clustering, dimensionality reduction;
- Text mining, NLP;
- Chatbot development;
- Application development, containers, docker
- IBM Cloud, Kubernetes;
- Node Red

First main objective of the project is to present a tool that is capable to generate an insight regarding the personality traits a person, given their writing example as an input.

Second main objective is to develop a chatbot that is capable of making database queries based on user's input and return a corresponding result.

The Roadmap

The roadmap consists of a set of subtasks and subproblems of the project, that have to be furnished to fulfill the principal objective, or that might yield better results. One must have in mind that some of them might change, be replaced or removed during the development of the project. Nonetheless, a roadmap serves as a guiding light at early stages of its evolution. For the moment, three principal branches of the development of project may be defined:

I. Identification of personality profiles

II. Chatbot development

III. Cloud Infrastructure

Subtask	Branch of the project	Estimated workload	Deadline	Completion
Chatbot prototype	Chatbot	25	End of June	1
Text Data mining	Personality	50	End of May	1
Exploration and learning of IBM Cloud	Cloud	32	End of May	1
Enterprise introduction and current	-	14	Mid June	1
Text Data cleansing	Personality	20	Start of June	1
Personality Insight	Personality	20	Start of June	1
Chatbot structure	Chatbot	16	Mid July	0.8
Entities, Intentions, Dialogs	Chatbot	24	End of July	0.8
Study of clustering methods	Personality	80	Start of August	0.75

Subtask	Branch of the project	Estimated workload	Deadline	Completion
Study of dimensionality reduction methods	Personality	42	End of July	1
Study of clustering evaluation techniques	Personality	32	Mid July	1
Data Visualization	Personality	24	End of August	0.9
Personality App	Personality	64	Start of September	0.2
App containerizati	Cloud	24	Mid September	0.5
Chatbot deployment	Chatbot	24	Mid September	0
Chatbot-Database connection	Chatbot	48	Start of September	0.2
Cloud CLI and Kubernetes	Cloud	36	End of August	0.1

The workload estimations are, of course, approximate values of time needed to fulfill the objective. Some of those are expected time estimations, while others are approximate amounts of time spent for already finished tasks.

Completion column corresponds to the share of necessary work that is already done.

1. Context, objectives and evolution of the internship

Data is the fuel that drives both research and business to innovation. It creates new opportunities and even new domains for the first and opens previously unseen possibilities for the second. And very often the process of innovation is what creates a common ground for the two: business struggles to achieve competitive edge, while research always struggles for funding.

Thus, all the major companies across all the industries have been creating their "Research & Development" departments with a goal to gain an advantage over the competitors and to improve their own products and services. The given internship is being carried out in the R&D department of housing supplier company "Vivialys Groupe".

In the circumstances of the uncertainty and dynamic development of the modern world, it is nothing but reasonable to expect any project to undergo changes and to evolve. It is important to cherish the original idea, but also to keep the final purpose and to react to the changing environment based on the new available information.

Initially, the main objective of the internship was to develop with the use of artificial intelligence a smart chatbot, that is capable of detecting human's personality. The bot itself should offer a common functional of presenting an information, such as information about the company or its programs based on user's entry. Besides that, other key feature of the chatbot is the ability to perform queries to the available apartments database following user's criteria. All in all, first concept of the chatbot was meant to have following functional:

- Basic communication, information transmission;
- Apartments database querying;
- Personality identification.

However, with the passage of time it became clear that those functionalities are not compatible among them. The very existence of chatbots revolves around standardization processes: to successfully mimic a conversation, all of the bot's

questions should be defined so, that any possible response from a user is considered and expected, so it can be treated correctly.

On the other hand, personality identification loses meaning when applied to homogeneous, standardized instances. It is a challenging endeavor that demands significant amounts of data during training phase, as well as it has a strict requirements for the input in the production. It is said, that, theoretically, machine learning algorithm may be trained to perform any task a human can, with sufficient data and well-thought approach to the training, of course. So, it is logical to ruminate on the amount of data that a human being needs to, say, classify other humans into extraverts and introverts. What if the goal is to also consider other human's needs, values and various traits of personality ?

The difficulty of the task and the data requirements scale up very quickly. Certainly this requirement exceeds few dozens of words, that one can obtain with the help of chatbot. This is why, the initial objective was divided into two different, though connected projects: a smart chatbot that would help users to navigate the site and make the queries to the database and a stand alone tool for personality identification.

It is worth to shed some light on the methodology of personality identification.

First of all, there are two fundamental theories, which any personality study is based upon. Those are:

- Lexical hypothesis, which is defined by two postulates. The first states that those personality characteristics that are important to a group of people will eventually become a part of that group's language. The second follows from the first, stating that more important personality characteristics are more likely to be encoded into language as a single word [1].

- Linguistic relativity, which is a hypothesis claiming that the structure of a language affects its speakers' world view or cognition, and thus person's perceptions are relative to their spoken or written language [2];

Although both of those are subject to critique, they are essential to the very concept of personality classification, as in their weakest forms the theories basically only state that the language and cognition are interconnected — the type of relation is what separates them and also what fuels the argument in the field.

In the scope of the current study the idea of personality identification consists in generating a particular insight regarding person's personality from an example of their writing. This insight is intended to be represented by a personality class prediction, some recommendations on the communication with this person and any other possibly relevant information.

There exist a few personality classification systems, such as big5 [3], 4 colors [4], Myers-Briggs type indicator [5]. The latest is considered to be the most applicable at the most, however, there is no unanimous decision on which systems describes existing types of personalities best. There is even no agreement on the fact that such system can be created in a way to describe all the possible variations in human behavior.

This question bothered even ancient minds. For instance, Hippocrates reflected upon four types of temperament — sanguine, phlegmatic, choleric, and melancholic. Significant progress on the matter was made in 20th century with the development and availability of the statistical methods. Nonetheless, the polémique in the field is still vivid and no definite conclusion has been drawn.

On the technical side, to fulfill the aforementioned objectives it was decided to use a vast functionality of IBM Cloud. Cloud development is arguably the most efficient approach, especially, when it comes to data science and digital transformation, that is available today.

Same as Google Cloud or Amazon Web Services, IBM Cloud offers to develop and deploy machine learning models, create data pipelines and data dashboards, organize clusters and deploy applications on them, etc. Besides that, IBM Cloud has a number of high level services, such as Personality Insight, Watson Assistant or Watson Discovery, that are capable to accelerate the development process.

2. Identification of personality profiles

1.1 Data

Keeping in mind the objective of the project, one should begin with its most important element — data.

As the goal is to produce an insight from an example of person's writing, the data one is going to be working on is text. There are no labels available at our disposal, which means that it is unknown which personality type each writing instance belongs to. Furthermore, it is unknown how many such types there might be. One may surely consider already existing studies on the matter, but none of them should be taken in a belief kind of way.

Thus, a given task should be treated as an unsupervised learning one, namely clustering, due to the aim is to define groups with similar characteristics in an unlabeled dataset. Machine learning and AI-based algorithms need a lot of data to be trained on. Two sources that will provide this data are enterprise clients mails database and an automated topic-based Twitter extraction.

First one speaks for itself — it is a set of plain text writings of company's clients. However, this database is not nearly enough to fulfill our needs, that is why it is reinforced with a second source. An example of script downloading the tweets of users related to the given topic is depicted in Fig. 2.1.

```
Searchig for users on 'Bukowski'
<<<>>>
Gathered 170 users and their timelines. Moving on.
Tweets of 159 users satisfy the given constraints. They are succesfully processed and saved.
<<<>>>
Total users gathered:  576
REQUESTS_MADE:  1058

Searchig for users on 'jazz legends'
<<<>>>
Gathered 15 users and their timelines. Moving on.
Tweets of 10 users satisfy the given constraints. They are succesfully processed and saved.
<<<>>>
Total users gathered:  586
REQUESTS_MADE:  2266
```

Figure 2.1 Twitter extraction script

It is worth to mention, what is being downloaded are not separate tweets, but whole timelines of a user, who was detected to tweet anything on the given topic. In such a way, a reasonable amount of one person's writing is obtained to be considered

representative. The timelines are processed using regular expressions to clean them from emojis, symbolic emojis and irrelevant symbols, such as "@" before names of other Twitter users.

The topics span from very general to specific, France or housing related ones. It was decided to consider profiles as much diverse as possible to increase the generalization power of ML models, trained on this data.

Although one would like to determine personality type of a person given their writing example, text data is not what is used as a direct input to a model. Before feeding the data to the model it passes through a very important step of a pipeline — Personality Insight Service of IBM Cloud. Its principle of work will be described thoroughly in section 4, but it is necessary to be introduce at this moment for the for the sake of congruency and simply to follow an order.

This service allows one to use a machine learning model that was trained on an enormous dataset (more than a million users for English) to produce a personality profile of a person from a text input. It has three groups of parameters:

- Big5
- Values
- Needs

Those three groups in total comprise 22 following parameters:

Big5	Values	Needs I	Needs II
Openness	Conservation	Challenge	Liberty
Conscientiousness	Openness to change	Closeness	Love
Extraversion	Hedonism	Curiosity	Practicality
Agreeableness	Self-enhancement	Excitement	Self-expression
Emotional range	Self-transcendence	Harmony	Stability
-	-	Ideal	Structure

Table 2.1 — Personality traits

These parameters describe various traits of human personalities and they take their values in the [0,1] interval. A fragment of the obtained profiles dataset may be found in Fig.2.2.

	UID	Openness	Conscientiousness	Extraversion	Agreeableness	Emotional range	Challenge	Closeness	Curiosity	Excitement	...	Love
0	89168924	0.672358	0.515518	0.202560	0.388493	0.952090	0.449133	0.491036	0.373575	0.361041	...	0.669498
1	74580436	0.769577	0.442860	0.063138	0.152624	0.998766	0.792551	0.712586	0.726484	0.940915	...	0.964348
2	52536879	0.445128	0.776801	0.899640	0.881138	0.706536	0.857511	0.842577	0.751038	0.759625	...	0.899586
3	17243213	0.707812	0.725007	0.057476	0.296220	0.990815	0.551724	0.447475	0.519926	0.487398	...	0.542529
4	278662460	0.659059	0.788241	0.708605	0.238162	0.566861	0.686579	0.331890	0.248036	0.528839	...	0.247863

Fig. 2.2 — Personality profiles

It should be noted, that it would have been possible to train the algorithms directly on text. However, in this way one would need to adapt a more sophisticated preprocessing approach, use embeddings to transform each word into a vector of an arbitrary size and also define some functional layers, in case of ANN, or transformations, otherwise. Moreover, at the final stage it is quite likely that the data would be flattened into a vector of floats, again, of an arbitrary size. Such approach is very demanding in terms of both dataset size and training time.

On the other hand, a possibility to use Personality Insight, which produces a result that fits our purpose, and was trained with labels, obtained from countless social media surveys, is very promising.

All in all, this tool is too powerful to be neglected in the scope of the project.

1.2 Unsupervised learning and clustering

1.3 Dimensionality reduction

1.4 Evaluation of the partition

3. Chatbot development

2.1 The main idea and description of bot's desired capabilities

2.2 Dialog structure, entities and intentions

4. Cloud Infrastructure

3.1 IBM Cloud

3.2 Databases

3.3 Docker, Kubernetes & Containers

Conclusion

Bibliography

1. Crowne, D. P. (2007). "Personality Theory", Don Mills, ON, Canada: Oxford University Press.

2. Hoijer, Harry, (1954). "Language in culture: Conference on the interrelations of language and other aspects of culture", Chicago: University of Chicago Press

3. Goldberg L.R. (1981). "Language and individual differences: The search for universals in personality lexicons". Review of Personality and social psychology. 1. Beverly Hills, CA: Sage.

4. Hartman, Taylor (1998). "The color code", New York : Scribner.

5. Myers, Isabel Briggs with Peter B. Myers (1980). "Understanding Personality Type", Mountain View, CA: Davies-Black Publishing.

6.

7.

Annex A